PhD Study with the Humanising Machine Intelligence Project at ANU

(Philosophy, Computer Science, Sociology, Political Science, Law)

In advanced industrial societies, we are increasingly dependent on algorithmic systems built around data and AI. These systems are reshaping the welfare state and the administration of criminal justice . They are used to police tax evasion , track down child abusers , and model the path of a pandemic . And they are used to weaponize vast surveillance networks through facial recognition technology . But algorithmic power extends far beyond the state. We spend ever more time working, socialising, and consuming within tech platforms. Our experiences are governed by algorithms that are constantly monitoring and shaping our behaviour and our attention, automatically selecting what we see, and what we don't see. These online experiences have offline consequences, among them an unprecedented challenge to democratic institutions worldwide. At the same time, tech companies and governments alike are investing billions in developing the infrastructure and research for the next major advance in AI, the next essential platform, the next once-human service that we can automate. We are, in many ways, in the middle of the most exciting, and the most dangerous, technological revolution in recent history.

The Humanising Machine Intelligence project at ANU was established in 2019 to bring together some of the university's world-leading researchers to understand, design, and develop democratically, constitutionally, and culturally legitimate data and AI systems. Our team comprises ten chief investigators from philosophy, computer science, political science, sociology and law, as well as nine research fellows and thus far three PhD students. Our approach is to bring the best of each of our constituent disciplines together to tackle common problems: first understanding the risks and opportunities associated with existing data and AI systems, then answering the foundational questions that must be addressed to build our shared values into machine systems, before designing both technologies and social structures that realise those values.

We invite applications to join this research community through the ANU PhD programs in computer science, philosophy, political science and international relations, law or sociology. HMI PhD students will develop expertise in their home discipline, but will also build the wider set of skills necessary to advance democratically legitimate machine intelligence. An HMI PhD will provide the disciplinary foundations necessary to conduct cutting-edge research at the world's leading research institutions—whether they be in academia or in industry.

Our project is deeply collaborative, with weekly research meetings, alternating between the Data, AI and Society public seminar and work-in-progress meetings of the working groups on our four core themes. Despite the pandemic, we remain a fundamentally international project, with ongoing collaborations with Stanford, Harvard, Carnegie Mellon, Princeton and Cambridge among others.

We seek candidates with deep foundations in at least one of the cognate disciplines underpinning the project, with an enthusiasm for learning from other fields, and a commitment to ensuring that the AI systems that we develop advance values that we all endorse.

We are interested in both Domestic and International Students. The deadline for International Students is 31 August 2020, though there will be another round in April 2021. For Domestic Students the deadline is October 31, with another round in mid-April.

Research Themes

Automating Governance/Governing Automation

This research theme is about the exercise of power through data and AI systems, and about how it can be rendered both just and legitimate. Power is used justly if is used to do the right things—but it is legitimate when it's also appropriately responsive to the will of the governed. Our research on this theme will include empirically-grounded analysis of precisely how data and AI are being used in the exercise of power—for example, analysis of whether algorithmic decision-making in the public sector is consistent with existing public law, or the moral critique of face surveillance. We'll explore the different approaches to data protection regulation in China, Australia and Europe, and look at the exercise of power through data and AI in the global South.

We'll also address foundational questions that help us figure out how this power can be used justly and legitimately—including fundamental theoretical work on the moral and political importance of explanations for decisions involving the exercise of power, as well as work in theoretical AI on how machine learning and planning algorithms can be rendered explicable. We'll be bringing theoretical AI, political philosophy, and law together to think through what it means for data and AI to be used fairly, and to design systems that do not wrongfully discriminate.

And we'll be putting those answers to foundational questions into practice, working with partners to advance the just and legitimate use of data and AI in the exercise of state and platform power. This will mean being a vocal contributor to policy debates within Australia, developing model regulation, working with the university to ensure that our own approach to data protection is justified, and working on specific projects aiming to use data and AI for public benefit in just and legitimate ways.

Personalisation

One of AI's most commercially successful applications—and one of the drivers of innovation—has been to use data and inferences about you to provide a personalised user experience: product recommendations, micro-targeted adverts, tailored newsfeeds, tailored prices. This automated personalisation has one goal: to affect your behaviour, in the pursuit of either profit or power. The means to that goal might be delighting you with a serendipitous recommendation, exploiting your cognitive biases, or even just holding your attention.

Personalisation is responsible for many of the greatest successes of data and AI, but also for many of the ways in which our social and political worlds are changing. Personalised news feeds create filter bubbles. Personalised social media inspire echo chambers. Personalised prices and ads can get you what you want just when you want it, but can also facilitate discrimination. Personalisation also involves chunking us up into unusual new social groups, as machine learning algorithms make inferences based on our behaviour. And personalisation creates new kinds of social goods that can be fairly or unfairly distributed—such as the basic good of attention.

In the discovery stream of this subproject, we are exploring online behavioural advertising from legal and political perspectives, asking whether existing consumer protection law is up to the task, as well as addressing the broader political question of whether personalised advertising and broader 'dark patterns' that influence us, often without our knowing they are doing so, should be considered morally objectionable at all. We have conducted research into how the YouTube recommender system works, and in particular how it funnels people towards more extreme political views, as well as work in computational social science on the allocation of attention by major online platforms. We are developing research projects that look at how and whether filter bubbles form on reddit, and when diverse participants reach consensus.

There's no doubt that personalisation has benefits as well as costs. The foundational questions raised here are how to minimise the downsides, not how to eliminate personalisation entirely. How can we design recommender systems that give us the serendipity without undermining our privacy, or creating new, illegitimate sources of power? Given that attention is going to be distributed algorithmically—and that it has real social consequences for those who get it or don't—how can we ensure that attention is distributed fairly? How should we rethink the ideal of free speech in a time of near costless communication? If you get all your information from a particular tech platform, how can they exercise that power responsibly, in a way that advances rather than undermines democracy? What questions for social theory and philosophy of science are raised by these new forms of market segmentation?

The goal of our project is to first figure out what the problem is, then answer the foundational questions that must be answered to make progress, and then put our answers to those questions into practice by designing better socio-technical systems. This will mean working with partners (especially in the private sector) to determine, for example, whether and how insurance prices can be personalised using data and AI, and how to deliver personalised opportunities in ways that reflect an underlying commitment to fairness.

Algorithmic Ethics

As well as using data and AI to support the delivery of various government and consumer services, we're actively developing AI-based systems that can effect significant (and not always predictable) state-changes without human intervention. To do this, when the moral stakes are high, we need to design moral

(legal, social etc) considerations into those systems. But how can we do this? In the other subprojects on governance and personalisation, we are exploring how to deliver social results with data and AI that better instantiate and promote our values. But are there any general technical solutions to the problem of how to design our values into AI systems? When those systems are acting autonomously—in the sense that they can make morally significant state-changes without the intervening decision of a human supervisor—how can we ensure that they robustly choose for the better?

In the discovery phase of this subproject, we're exploring different kinds of automation, determining where a successful integration of moral reasons is most needed, and where it is most likely to be achievable. We are exploring robotic systems, autonomous vehicles, planning algorithms, and the more general use of reinforcement learning.

The foundational questions at issue are especially challenging. There are numerous first-order questions in moral philosophy that we must answer, such as how to think about moral decision-making under risk, and over time—there is at present no sequential moral decision theory to provide an analogue to the sequential decision theory used in AI. We even need a metaethics for ethical autonomous systems—how do we decide how to decide which values to incorporate? Are we aiming to approximate moral truth, or to give due weight to disagreement and uncertainty? Should we hold autonomous systems to the same standards as humans in the same decision situations, or to higher ones? How can we formalise these moral constraints and objectives in ways that are interoperable with planning and machine learning languages? What are the complexity constraints on making progress?

On the design side, our key goal is to implement some of these answers to foundational questions in actually operational autonomous AI systems. We are exploring partnerships in robotics and with autonomous vehicle companies. As well as designing more ethical autonomous systems, we are also advancing research on how to verify and assure ourselves of their optimal ethical performance.

Human-AI Interaction

If we designed AI systems that were morally perfect in a vacuum, but didn't take into account the predictable way people react when interacting and using those systems, then we would end up with very bad AI systems. Our conception of machine ethics and our design of interactive systems should aim to minimise harms and optimise goods in machines, ourselves and the outcomes of our interactions. This has two sides.

First, it's no good designing systems that would deliver optimal outcomes if they were used only by perfectly rational human operators. People have predictable cognitive habits and biases, as well as other prejudices and moral failures. If these aren't taken into account, then we will reliably end up aggravating harms and introducing injustices.

Second, we need to keep in mind the dynamic effects of human-AI interaction, most importantly the ways in which how we work with automated systems will change us. Our goal should be to design automated systems that help us act better, morally speaking, not to design systems that make better decisions than we would, on our behalf.

The first stage of research in this theme is identifying the human attributes that need to be taken into account when designing automated systems. There is, of course, a substantial field of work on human computer interaction, and 'human factors', which we will be engaging with—though our combination of philosophy, sociology, and social psychology will bring something new to that field. We're particularly interested in exploring the cognitive defects that AI systems must account for (and might exacerbate), as well as the human skills—including skilled moral behaviour—that automation might ultimately undermine.

The second phase is to determine what kind of interactions between humans and AI systems we want to aim for, what the goals and constraints are. For example, is there such a thing as moral skill, which can be weakened or strengthened by increased reliance on machines? In what other ways are our technologically mediated lives changing us, and how can we direct those changes for the better?

The design phase of this subproject will see us determine how we can shape machines so that they help make us better people, rather than just making better decisions on our behalf. It also means anticipating the kinds of cognitive errors that we are likely to make, as users of AI, and mitigating them in advance—for example by representing the actual degree of uncertainty in a choice more clearly, or by not just doing the right thing, but providing assurance that the right thing is being done.

Why ANU and HMI

The Australian National University is Australia's leading university for philosophy, the social sciences, and theoretical AI. Philosophy at ANU is ranked seventh in the world (QS 2020), with particular strengths in political philosophy (2nd, Philosophical Gourmet [PGR]), and the philosophy of science (for example, 5th in decision theory, 10th in philosophy of cognitive science, all PGR). ANU Political Science and International Studies, and Sociology, are also in the global top-10 (per QS). Computer Science at ANU ranks 40th in the world, and for the areas of computer vision, machine learning and data mining, and AI, ranks top in Australia (CSRankings.org).

Because of its reputation, ANU is a hub for attracting the world's leading researchers. HMI has a global network of internationally respected experts in each of our areas of study, who regularly visit the ANU, and collaborate on workshops and conferences both locally and internationally.

The HMI project is a deeply interdisciplinary and collaborative project with an outstanding team of faculty. We already have partnerships with industry, government and other research organisations in the UK, US, Europe, and China. Our PhD students will have the opportunity to collaborate with scholars in universities such as Carnegie Mellon, MIT, Oxford, Northeastern, Stockholm, Cambridge, and Stanford.

HMI PhD students will be housed in their home discipline, and will have a member of our team in their home discipline as chair of their panel, but they will have team-members from at least one other discipline on their panel, and they will be encouraged both to master their own methodology, and to develop deep competence in the other core areas of the project.

Requirements

• A high first class honours degree or equivalent in a relevant discipline

• Fluent oral and written communication in English

• Passion and motivation for position

• Track record or potential for academic excellence in their chosen field

• History of or enthusiasm for engaging with and contributing to other fields

• Other requirements as stipulated by the receiving discipline

Information

ANU has a range of scholarships available to both local and international PhD students. For more information please visit the scholarships page.

Application

If interested, you should reach out to one of the CIs on the HMI project in the first instance. They can then guide and support you through the application process, including the development and refinement of suitable projects. We are very keen to develop these collaboratively, so if you have an idea that you think fits within the remit of HMI, please reach out.